マルコフ決定過程

⑴エージェントが環境の中で行動を選び、報酬を得ながら最適な戦略を学ぶモデル

⑵未来は現在の状態と行動だけで決まる（マルコフ性）

⑶「累積報酬が最大になるような戦略（方策）」を見つけるのがゴール。

（式） $\rho(s',r|s,a)=P\{St+1=s',Rt+1=r|St=s,At=a\}$

$\rho(s',r\|s,a)$	状態sで行動aをとったとき、次の状態がs'になり、報酬がrになる確率
P{}	確率関数
St	時刻tにおける状態
At	時刻tにおける行動
St+1=s′	次の状態がS'になる事象
Rt+1=r	次の状態の報酬がrになる事象

（状態Sで行動aをとったとき、次の状態がS'になり、報酬がrになる確率） =
　時刻tにおける状態がsで、時刻tにおける行動がaのとき、状態がS'になり報酬rを受け取る確率

⇓

状態s_tで行動aをとったとき、次の状態がs'になり、報酬がRt+1=rとなる確率

状態遷移確率

p(s'|s,a) =$\sum_{r} p(s', r|s, a)$

p(s'|s,a)	状態sで行動aをとったとき、次の状態がs'になる確率
$\sum_{r} p(s', r\|s, a)$	状態sで行動aをとったとき、次の状態がs'になる確率の総和

状態sで行動aをとったとき、次の状態からs'になり、報酬がrになる確率＝状態sで行動aをとったとき、次の状態がs'になる確率の総和

（例）
状況
・現在の状態：s_t=s(プレイヤーが部屋Aにいる)
・行動：A_t=a(プレイヤーが右に進む)
・次の状態：s_{t+1}=s'(プレイヤーが部屋Bに移動)
遷移と報酬の確率
・部屋Bに移動して10ポイントもらう確率=0.3
・部屋Bに移動して5ポイントもらう確率=0.5
・部屋Bに移動して0ポイントもらう確率=0.2

→部屋Bに移動する場合でも、報酬がいくつかの異なる値をとる可能性がある場合の状態s'に遷移する確率を求める

p(s'|s,a)=0.3+0.5+0.2 = 1.0
つまり、部屋Bに移動する確率は100%

マルコフ決定過程の報酬の期待値

r(s,a,s')=$\sum_{r} r \cdot p(r|s,a,s')$

r(s,a,s')	状態sで行動aをとり、次の状態s'になったときの期待報酬
r	得られる報酬の値
p(r|s,a,s')	状態sで行動aをとり、次の状態がs'になったときに、報酬がrになる確率
$\sum_{r}$	全ての報酬rにわたって合計

状態sで行動aをとり、次の状態s'になったときの期待報酬=全ての報酬rにわたって合計(得られる報酬の値×状態sで行動aをとり、次の状態がs'になったときに、報酬がrになる確率)

(例)
状況
・現在の状態s:プレイヤーが「部屋A」にいる。
・行動a:「右に進む」
・次の状態s':「部屋B」に移動

報酬の確率
・部屋Bに移動して報酬+10 = 0.3
・部屋Bに移動して報酬+5=0.5
・部屋Bに移動して報酬+0=0.2

報酬の期待値を計算

10×0.3+5×0.5+0×0.2=5.5

→プレイヤーが部屋Aで「右に進む」という行動をとり、次の状態(部屋B)に移動したとき、期待される報酬の平均は5.5となる。

割引累積報酬

⑴エージェントが行動した結果得られる将来の報酬の合計を求めたもの。

・すでに取った行動と得られる報酬が分かっている特定のエピソード（試行）で、時刻tの行動の割引累積報酬を求める式

$$G_t = \sum_{k=1}^{\infty} r^{k-1} R_{t+k}$$

G_t	時刻tからの割引累積報酬
R_{t+k}	時刻t+kで得られる報酬
$\sum_{k=1}^{\infty}$	将来の報酬をすべて合計
r	割引率
r^{k-1}	報酬を時刻が進むごとに減衰させる(時間の重み付け)

時刻tからの割引累積報酬=時刻tからの将来の報酬をすべて合計(時間の重み付け×時刻t+kで得られる報酬)